Bayesian varying coefficient models using PC priors

Bayesian varying coefficient models using PC priors

Maria Franco-Villoria
Department of Economics and Statistics
University of Torino

Massimo Ventrucci
Department of Statistical Sciences
University of Bologna

Håvard Rue
CEMSE Division, King Abdullah University of Science and Technology,
Thuwal, Saudi Arabia

June 7, 2018

Abstract

Varying coefficient models arise naturally as a flexible extension of a simpler model where the effect of the covariate is constant. In this work, we present varying coefficient models in a unified way using the recently proposed framework of penalized complexity (PC) priors to build priors that allow proper shrinkage to the simpler model, avoiding overfitting. We illustrate their application in two spatial examples where varying coefficient models are relevant.

KEYWORDS: varying coefficient models; penalized complexity prior; INLA; overfitting

ADDRESS FOR CORRESPONDENCE: Maria Franco-Villoria, Department of Economics and Statistics Cognetti de Martiis, University of Torino. Email: maria.francovilloria@unito.it

1 Introduction

Varying coefficient models (Hastie and Tibshirani, 1993) can be seen as a general class of models that encompasses a large number of statistical models as special cases: the generalized linear model, generalized additive models, dynamic generalized linear models or even the more recent functional linear models. In practice, varying coefficient models (VCMs) are useful in presence of an *effect modifier*, a variable that "changes" the effect of a covariate of interest on the response.

VCMs arise in a vast range of applications, including economics (Gelfand et al., 2003), nutrition (Hoover et al., 1998), ecology (Ferguson et al., 2007; Finley, 2011), air quality (Mu et al., 2018), epidemiology (Fan and Zhang, 1999) and survival analysis (Cai and Sun, 2003; Tian et al., 2005). The most commonly used effect modifiers are time (Hoover et al., 1998; Fan and Zhang, 1999; Ferguson et al., 2007) and space (Gelfand et al., 2003; Finley, 2011; Mu et al., 2018) but other variables can also be considered (Hastie and Tibshirani, 1993). Parameter estimation for this kind of models has been approached both in a frequentist and Bayesian framework. In a frequentist setting, a varying coefficient is usually considered as a smooth function that can be estimated using a kernel smoother (Fan and Zhang, 1999; Park et al., 2015) or a linear combination of basis functions such as splines (Hastie and Tibshirani, 1993; Huang et al., 2002; Marx, 2010); for a comprehensive review

on estimation procedures see Fan and Zhang (2008). In a Bayesian setting, the usual approach is to describe the varying coefficient by a vector of random effects distributed at prior as a Gaussian Markov Random Field (GMRF), see Rue and Held (2005).

For the sake of a general notation that includes all cases discussed in this paper, consider the triplet (y_t, x_t, z_t), $t = 1, ..., n$, observed on n observational units, with z being the variable modifying the relationship between the covariate x and the response y. The effect modifier can either be a continuous variable (e.g. temperature) or a time/space index (e.g. day or municipality). Assuming y belonging to the exponential family, the linear predictor of a generalized VCM is

$$\eta_t = \alpha + \beta(z_t)x_t \qquad t = 1, ..., n,$$

where $\beta(z_t)$, $t = 1, ..., n$, is the varying regression coefficient (VC), that can be regarded as a stochastic process on the effect modifier domain. For ease of notation we will use β_t to denote $\beta(z_t)$.

While the flexibility that VCMs offer can be desirable in certain applications and much work has been devoted to the development of flexible models, we should keep in mind that flexibility is a relative concept. It is natural to think about a varying coefficient model as a flexible extension of a simpler model; for example, we can consider increasing the flexibility of the simple linear regression model $\eta_t = \alpha + \beta x_t$, $t = 1, \ldots, n$ by allowing the coefficient β to vary over t.

In a Bayesian context, we can envision several models for $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)^T$, depending on what we think the structure of the VC to be in the application at hand. For instance, we can assume exchangeability over $t = 1, \ldots, n$ with $cor(\beta_i, \beta_j) = \xi$ for $i \neq j$ if there is no natural ordering among the values of z. If the effect modifier is time, β_t might be a 1^{st} order autoregressive (AR1) and ξ the lag-one correlation, or a spline if we want to ensure smoothness. The coefficients may also vary in space in a continuous or discrete way, in which case a Gaussian random field with a certain covariance function or a conditionally autoregressive (CAR) model can be assumed, respectively. Even though these models have been treated separately in literature and might look like different models at first, they can all be gathered under a unified framework where the coefficients are defined by a Gaussian process.

In a fully Bayesian framework, a prior distribution has to be specified for the hyperparameter(s) of the Gaussian process. Common choices of the prior might lead to overfitting, i.e. might *push* the model away from the simpler model even when a more flexible one is not appropriate (Frühwirth-Schnatter and Wagner, 2010, 2011). Even though we might have solid scientific motivation to believe that a varying coefficient is needed, we should bear in mind that the VCM rises naturally from a simpler model. Ensuring that the prior gives a chance to the simpler model becomes then fundamental in a varying coefficient model, allowing to deviate from it only if there is evidence in the data for doing so.

The goal of this paper is twofold: to present varying coefficient models in a unified way and to build priors for these models that allow proper shrinkage to the simpler model, avoiding overfitting. For doing so, we use the "Penalized Complexity (PC) Prior" framework (Simpson et al., 2017), where a model component is considered as a flexible extension of a simpler version of the model component, referred to as the base model. PC priors are defined on the scale of the distance from a base model and then transferred to the scale of the original parameter by a standard change of variable transformation. This strategy can be applied to different models for $\boldsymbol{\beta}$ describing the VC in a unique way, as the base model can always be easily identified in terms of a value for ξ. In this sense, PC priors represent a unified framework to build the prior in a VCM setting. We derive PC priors for various varying coefficient models and illustrate their implementation in real case studies. Even though the mathematical expression for the PC prior under the different models for $\boldsymbol{\beta}$ might look different, the prior is always built in the same way using well defined principles.

The plan of the paper is as follows. Section 2 presents varying coefficient models in a unified way, while the general framework to construct PC priors is briefly reviewed in Section 3. In Section 4, several PC priors for ξ are derived under different model choices for $\boldsymbol{\beta}$, focusing first on the unstructured case (Section 4.1), where the realizations of the VC are assumed to be exchangeable. Structured cases, such as time and space are presented in Sections 4.2 and 4.3. Examples are illustrated in Section 5. The paper closes with a discussion in Section 6.

2 Varying coefficient models

Let us now specify a Bayesian VCM, seeing it as a flexible extension of the simple linear regression model $\eta_t = \alpha + \beta_0 x_t$, which will be denoted as *base model*; this can be thought of as the fit obtained if data do not show evidence for a varying coefficient but for a constant regression coefficient instead. Without loss of generality, we can assign the prior $\beta_0 \sim \mathcal{N}(0, 1)$ to the base model:

$$\begin{array}{ll} \eta_t = \alpha + \beta_0 x_t & t = 1, ..., n, \\ \beta_0 \sim N(0, 1). & \end{array}$$

If we believe that the covariate effect is not constant in z, we can allow for deviation from β_0 in the form of a varying coefficient model,

$$\begin{array}{ll} \eta_t = \alpha + (\beta_0 + \beta_t)\, x_t & t = 1, ..., n, \\ \boldsymbol{\beta}|\xi \sim \pi(\boldsymbol{\beta}|\xi), & \end{array} \tag{1}$$

where $\boldsymbol{\beta} = (\beta_1, \ldots, \beta_n)^T$ is a vector of random effects defining a stochastic process over z, denoted as $\pi(\boldsymbol{\beta}|\xi)$ with ξ the associated hyper-parameter(s).

In what follows we will assume the linear predictor $\eta_t = \alpha + (\beta_0 + \beta_t)\, x_t$ in Equation (1) and consider different Gaussian models for $\pi(\boldsymbol{\beta}|\xi)$.

2.1 The unstructured case

The simplest correlation structure for random effects is to assume that they are exchangeable; this is commonly used to account for dependence among repeated measures in longitudinal models (Laird and Ware, 1982). If $\boldsymbol{\beta} = (\beta_1, ..., \beta_n)^\intercal$ are exchangeable over $t = 1, \ldots, n$, then $\boldsymbol{\beta} \sim \mathcal{N}(0, \tau^{-1}\boldsymbol{R}(\tilde{\rho}))$ where

$$\boldsymbol{R}(\tilde{\rho}) = \begin{bmatrix} 1 & \tilde{\rho} & \ldots & & \tilde{\rho} \\ \tilde{\rho} & 1 & \tilde{\rho} & \ldots & \tilde{\rho} \\ \cdot & & \cdot & & \cdot \\ \cdot & & & \cdot & \cdot \\ \tilde{\rho} & \tilde{\rho} & \ldots & \tilde{\rho} & 1 \end{bmatrix} \tag{2}$$

and τ is a precision parameter. For $\boldsymbol{R}(\tilde{\rho})$ to be positive semi-definite, $-1/(n-1) \leq \tilde{\rho} \leq 1$ (Simpson et al., 2017). In the following, we consider $0 \leq \tilde{\rho} \leq 1$.

In this case, Model (1) can be reparametrized as $\eta_t = \alpha + \beta_t x_t$, $t = 1, \ldots, n$, with unit marginal variance and

$$\boldsymbol{\beta} \sim N(0, \boldsymbol{R}(\rho)). \tag{3}$$

A sensible base model is $\rho = 1$, corresponding to $\beta_t = \beta \ \ \forall t$.

2.2 The structured case: temporal variation

In many real life applications the values of the effect modifier follow a natural ordering, e.g. time, so that it is not realistic to assume exchangeability of β_t. Instead, autoregressive (AR) models from time series analysis can be adopted (Sørbye and Rue, 2017). An alternative is to consider the varying coefficient as a smooth function. A popular prior in the context of smoothing with splines is the 2^{nd} order random walk (RW2), that can be seen as a discrete representation of a continous (integrated) Wiener process that retains the Markov property and is computationally efficient (Lindgren and Rue, 2008). It is also used in P-splines (Marx, 2010) where a RW2 is assigned to the coefficients of local B-spline basis functions. In the following we consider three cases: the 1^{st} order autoregressive (AR1) and the 1^{st} and 2^{nd} order random walk (RW1, RW2). In all three cases, we always assume the linear predictor reported in Equation (1), but consider different models for β_i.

2.2.1 The autoregressive model of first order

The most common model for dependence on time is the autoregressive process of first order (AR1), the discrete-time analogue of the Ornstein-Uhlenbeck process, characterized by a correlation function with exponential decay rate. A 1^{st} order autoregressive prior on the varying coefficient is $\beta_t = \tilde{\rho}\beta_{t-1} + w_t$, where $|\tilde{\rho}| < 1$ represents the lag-one correlation, $w_t \sim \mathcal{N}(0, \tau^{-1}(1-\tilde{\rho}^2)), t = 2, ..., n$, and $\beta_1 \sim \mathcal{N}(0, \tau^{-1})$. The varying coefficient has a joint distribution given by $\boldsymbol{\beta} \sim \mathcal{N}\left(0, \tau^{-1}\boldsymbol{R}(\tilde{\rho})\right)$ with $\boldsymbol{R}(\tilde{\rho})_{ij} = (\tilde{\rho}^{|i-j|})$ and τ a precision parameter. Similarly to Section 2.1, we can reparametrize Model (1) as $\eta_t = \alpha + \beta_t x_t$, $t = 1, \ldots, n$, so that

$$\boldsymbol{\beta} \sim \mathcal{N}\left(0, \boldsymbol{R}(\rho)\right) \tag{4}$$

and $\beta_1 \sim N(0, 1)$. Also in this case the base model is $\rho = 1$, i.e. no change in time.

2.2.2 Random walk model of order one and two

We can consider the varying coefficient $\boldsymbol{\beta}$ in Equation (1) as a smooth stochastic process on the effect modifier scale. The equivalence between smoothing splines and Gaussian processes was shown in Kimeldorf and Wahba (1970). In a Bayesian framework, smoothing models are obtained using random walk priors on the varying coefficients. A random walk is an intrinsic Gaussian Markov Random Field (IGMRF, Rue and Held (2005) ch. 3), i.e. a process with the multivariate Gaussian density

$$\pi(\boldsymbol{\beta}|\tau) = (2\pi)^{-\mathrm{rank}(\boldsymbol{R})/2}(|\tau\boldsymbol{R}|^*)^{1/2}\exp\left\{-\frac{\tau}{2}\boldsymbol{\beta}^\mathsf{T}\boldsymbol{R}\boldsymbol{\beta}\right\} \tag{5}$$

where the structure matrix $\boldsymbol{R}$ is sparse and rank deficient ($\mathrm{rank}(\boldsymbol{R}) = n - r$), τ is a scalar precision parameter and $|\tau\boldsymbol{R}|^*$ is the generalized determinant. The precision parameter regulates the amount of shrinkage towards the base model, that corresponds to $\tau = \infty$. The structure matrix encodes the conditional dependencies among the coefficients $\boldsymbol{\beta}$, as $\boldsymbol{R} = \boldsymbol{D}_r^\mathsf{T}\boldsymbol{D}_r$, where $\boldsymbol{D}_r$ is a matrix such that $\boldsymbol{D}_r\boldsymbol{\beta} = \Delta^r\boldsymbol{\beta}$, with Δ^r the r^{th}-order difference operator. The structure $\boldsymbol{R}$ for a RW1 (Eq. (6)) and a RW2 (Eq. (7)) is:

$$\boldsymbol{R} = \kappa \begin{bmatrix} 1 & -1 & & & & & \\ -1 & 2 & -1 & & & & \\ & -1 & 2 & -1 & & & \\ & & \ddots & \ddots & \ddots & & \\ & & & -1 & 2 & -1 & \\ & & & & -1 & 1 \end{bmatrix} \quad (6) \qquad \boldsymbol{R} = \kappa \begin{bmatrix} 1 & -2 & 1 & & & & & & & \\ -2 & 5 & -4 & 1 & & & & & & \\ 1 & -4 & 6 & -4 & 1 & & & & & \\ & 1 & -4 & 6 & -4 & 1 & & & & \\ & & \ddots & \ddots & \ddots & \ddots & \ddots & & & \\ & & & 1 & -4 & 6 & -4 & 1 & & \\ & & & & 1 & -4 & 6 & -4 & 1 & \\ & & & & & 1 & -4 & 5 & -2 & \\ & & & & & & 1 & -2 & 1 \end{bmatrix} \quad (7)$$

where $\kappa \in \mathbb{R}$ is an appropriate scaling parameter, calculated as the geometric mean of the diagonal elements of the generalized inverse of $\boldsymbol{R}$, so that the marginal variance from the null space is equal to 1 (Sørbye and Rue, 2014). This rescaling is necessary to avoid scaling issues inherent in RW models, such as dependence on the graph (Sørbye and Rue, 2014).

The rank deficiency of the structure matrix also identifies the order r of the IGMRF. Model (5) describes deviation from a polynomial model of degree $r-1$): e.g. a constant for RW1 ($r=1$) and a linear trend for RW2 ($r=2$). This means we need to impose a sum to zero constraint on $\boldsymbol{\beta}$ to avoid confounding with β_0 in Equation (1), with the difference that using a RW2 will result in a smoother fit than if a RW1 is used.

An alternative base model for the RW2

As the RW2 describes deviation from a linear trend, a more natural parametrization of the varying coefficient considers $\alpha_0 + \beta_0 z_t$ as base model. In this case, we need to impose the constraint $\boldsymbol{A}\boldsymbol{\beta} = 0$ to ensure identifiability of both terms in the base model, where $\boldsymbol{A}_{2\,\mathrm{x}\,n} = [\mathbf{1}, \boldsymbol{l}]^{\mathsf{T}}$ and $\boldsymbol{l} = (1, \ldots, n)^{\mathsf{T}}$.

Without loss of generality, we assume equally spaced locations. The case of irregularly spaced locations differs only in the structure matrix $\boldsymbol{R}$ and the constraint matrix $\boldsymbol{A}$, that has to be modified with the inclusion of appropriate integration weights; see (Lindgren and Rue, 2008) for details.

2.3 The structured case: spatial variation

Spatially structured models include the cases of continuous or discrete spatial variation. In the former case, the effect modifier is the pair of (scaled) latitude and longitude coordinates, $\boldsymbol{z}_t = \{\mathrm{lat}_t, \mathrm{lon}_t\}$ and β_t can be assumed as a realization from a spatial process. The class of Gaussian Random Field (GRF) models equipped with a *Matèrn* covariance is the most popular model (Stein, 1999). For areal data, the spatial units are identified by a one-dimensional region index, with no unique ordering among the regions. Neighbouring regions are assumed to be correlated, and the neighbourhood structure can be coded into a structure matrix. To model β_t, the standard approach is to use conditionally autoregressive (CAR) models proposed by Besag (1974); see Waller et al. (2007); Staubach et al. (2002) for applications.

2.3.1 Areal spatial variation

Models for areal data have been widely discussed in the literature and are useful, for example, in epidemiological studies (Banerjee et al., 2015), where data are not available at individual level but only at some aggregated level such as municipality or zip code (see Figure 4 for an example).

Assume the linear predictor in (1) where $t = 1, \ldots, n$ indicates each of the non overlapping regions in a lattice. Areas i and j are considered as neighbours, denoted as $i \sim j$, if they share a

common border. The spatially varying coefficient $\boldsymbol{\beta} = (\beta_1, ..., \beta_n)^\mathsf{T}$ follows an Intrinsic Conditional Autoregressive (ICAR) model (Besag, 1974):

$$\beta_t|\boldsymbol{\beta}_{-t}, \tau \sim \mathcal{N}\left(\frac{1}{n_t}\sum_{j:t\sim j}\beta_j, (n_t\tau)^{-1}\right)$$

with n_t the number of neighbours of region t and τ a precision parameter. The base model, corresponding to no variation over area, is $\tau = \infty$. The joint distribution for $\boldsymbol{\beta}$ is

$$\pi(\boldsymbol{\beta}|\tau) = (2\pi)^{-(n-1)/2}(|\tau\boldsymbol{R}|^*)^{1/2}\exp\left\{-\frac{\tau}{2}\boldsymbol{\beta}^\mathsf{T}\boldsymbol{R}\boldsymbol{\beta}\right\} \tag{8}$$

where the structure matrix $\boldsymbol{R}$ is singular with null space $\mathbf{1}$ and entries:

$$R_{i,j} = \begin{cases} n_i & i = j \\ -1 & i \sim j \\ 0 & \text{otherwise.} \end{cases}$$

2.3.2 Continuous spatial variation

In this case, $t = (\text{lat}_t, \text{lon}_t)$, properly scaled, represents location within a spatial region $D \subseteq \mathbb{R}^2$ and the spatially varying coefficient can be seen as a realization of a Gaussian random field (GRF) with a *Matèrn* covariance function characterized by the marginal variance τ^{-1} and range parameter ϕ. These two parameters cannot be estimated consistently under infill asymptotics (Warnes and Ripley, 1987; Zhang, 2004), but only a function of those such as the product or the ratio, depending on the smoothness of the GRF.

Assuming the linear predictor in (1) the spatially varying coefficient

$$\boldsymbol{\beta} \sim \mathcal{N}\left(\mathbf{0}, \tau^{-1}\boldsymbol{R}(\phi)\right) \tag{9}$$

with $\boldsymbol{R}(\phi)_{ij} = (C(||i-j||))$, $C(\cdot)$ is a *Matèrn* correlation function with fixed smoothness ν:

$$C(h) = \frac{2^{1-\nu}}{\Gamma(\nu)}\left(\frac{\sqrt{8\nu}h}{\phi}\right)^\nu K_\nu\left(\frac{\sqrt{8\nu}h}{\phi}\right)$$

and h represents the distance between any pair of locations. The base model in this case corresponds to $\tau = \infty$, $\phi = \infty$.

The prior for all the hyperparameters τ (and ϕ) in Section 2 can be built in a unified way regardless of the assumed model for $\boldsymbol{\beta}$ using penalized complexity priors.

3 Review of Penalized Complexity (PC) Priors

In this section we summarize the four main principles underpinning the construction of PC priors, namely: support to Occam's razor (parsimony), penalisation of model complexity, constant rate penalisation and user-defined scaling. For a more detailed presentation of these principles the reader is referred to Simpson et al. (2017).

The PC prior framework offers a unified approach for constructing priors for all the various models considered in Section 2 while guaranteeing proper shrinkage to the base model. Note that even though the mathematical expressions reported in Section 4 look different, the PC prior is always the same for all models.

Let f_1 denote the density of a model component $\boldsymbol{w}$ where ξ is the parameter for which we need to specify a prior. The base model, corresponds to a fixed value of the parameter $\xi = \xi_0$ and is characterized by the density f_0.

1. The prior for ξ should give proper shrinkage to ξ_0 and decay with increasing complexity of f_1 in support of Occam's razor, ensuring parsimony; i.e. the simplest model is favoured unless there is evidence for a more flexible one.

2. The increased complexity of f_1 with respect to f_0 is measured using the Kullback-Leibler divergence (KLD, Kullback and Leibler, 1951),

$$\text{KLD}(f_1||f_0) = \int f_1(w) \log\left(\frac{f_1(w)}{f_0(w)}\right) dw,$$

which, for zero mean multivariate normal densities is

$$\text{KLD}(f_1||f_0) = \frac{1}{2}\left(\text{tr}(\boldsymbol{\Sigma}_0^{-1}\boldsymbol{\Sigma}_1) - n - \ln\left(\frac{|\boldsymbol{\Sigma}_1|}{|\boldsymbol{\Sigma}_0|}\right)\right)$$

where n is the dimension. For ease of interpretation, the KLD is transformed to a unidirectional distance measure

$$d(\xi) = d(f_1||f_0) = \sqrt{2\text{KLD}(f_1||f_0)} \tag{10}$$

that can be interpreted as the distance from the flexible model f_1 to the base model f_0.

3. The PC prior is defined as an exponential distribution on the distance, $\pi(d(\xi)) = \lambda \exp(-\lambda d(\xi))$, with rate $\lambda > 0$, ensuring constant rate penalization. Therefore, the mode of a PC prior is always at the base model. The PC prior for ξ follows by a change of variable transformation.

4. The user must select λ based on his prior knowledge on the parameter of interest (or an interpretable transformation of it $Q(\xi)$). This knowledge can be expressed in terms of a probability statement, e.g. $\mathbb{P}(Q(\xi) > U) = a$, where U is an upper bound for $Q(\xi)$ and a is a (generally small) probability.

PC priors follow specific principles that remain unchanged no matter the choice of the varying coefficient prior $\pi(\boldsymbol{\beta}|\xi)$. This means we can address prior specification for any varying coefficient model in the same way. Since PC priors are built on the distance scale and then transformed to priors on the parameter scale, they are invariant under reparametrization. One major advantage of these priors is that they prevent overfitting by construction, as they guarantee shrinkage towards the base model. PC priors for the marginal variance of a Gaussian random effect have been shown to outperform other priors widely used in literature (such as Inverse Gamma priors) when data are weakly informative or the size of the effects is close to the base model (Klein and Kneib, 2016). Finally, prior information, if available, can be coded into an intuitive way by simply specifying U and a.

4 PC priors for varying coefficient models

In this section we derive PC priors for the varying coefficient models discussed in Section 2. Within this framework, we can always build the prior for the corresponding parameter as an exponential distribution on the distance from the base model. Here we present the main results, while technical details can be found in the Appendix.

4.1 The unstructured case

As described in Section 2.1, the base model for Model (3) is $\rho = 1$. The PC prior for ρ:

$$\pi(\rho) = \frac{\theta \exp(-\theta\sqrt{1-\rho})}{2\sqrt{1-\rho}(1-\exp(-\theta))}, \quad 0 \leq \rho \leq 1, \quad \theta > 0. \tag{11}$$

The prior is scaled in terms of θ based on the prior belief that the user has about the parameter ρ in the form of (U, a) such that $\mathbb{P}(\rho > U) = a$. The corresponding value for θ is given by the solution of the equation

$$\frac{1-\exp(-\theta\sqrt{1-U})}{1-\exp(-\theta)} = a$$

that has to be solved numerically, provided that $a > \sqrt{1-U}$. The PC prior in (11) is illustrated in Figure 1.

4.2 The structured case: temporal variation

4.2.1 The autoregressive model of first order

For Model (4), Sørbye and Rue (2017) derive the PC prior with base model at $\rho = 1$ as

$$\pi(\rho) = \frac{\theta \exp(-\theta\sqrt{1-\rho})}{(1-\exp(-\sqrt{2}\theta))2\sqrt{1-\rho}}, \qquad |\rho| < 1, \quad \theta > 0. \tag{12}$$

The user can incorporate information on his/her prior belief about the size of the correlation parameter by setting U and a so that $\mathbb{P}(\rho > U) = a$. To work out θ the equation

$$\frac{1-\exp(-\theta\sqrt{1-U})}{1-\exp(-\sqrt{2}\theta)} = a, \quad a > \sqrt{(1-U)/2}$$

needs to be solved numerically for θ as in the unstructured case. The PC prior in (12) is illustrated in Figure 2.

4.2.2 Random walk model of order one and two

In the case of Model (5), the amount of deviation from the base model depends on τ, with base model at $\tau = \infty$. Simpson et al. (2017) derive the PC prior for τ as a Gumbel$(1/2, \theta)$ type 2 distribution

$$\pi(\tau) = \frac{\theta}{2}\tau^{-3/2}\exp\left(-\theta/\sqrt{\tau}\right), \quad \tau > 0, \theta > 0. \tag{13}$$

To derive the scaling parameter θ, Simpson et al. (2017) suggest to bound the marginal standard deviation, $1/\sqrt{\tau}$. This way it is sufficient to specify (U, a) and solve $Pr(1/\sqrt{\tau} > U) = a$ for θ, which gives $\theta = -\log(a)/U$.

To aid the user in specifying parameters (U, a), Simpson et al. (2017) provide a general rule of thumb: "setting $a = 0.01$, the marginal standard deviation of $\boldsymbol{\beta}$ with $\boldsymbol{R} = \boldsymbol{I}$ is about $0.31U$"; e.g. if we think a standard deviation of approximately 0.3 is a reasonable upper bound, we need to set $U = 0.3/0.31 = 0.968$. The PC prior in (13) is illustrated in Figure 3.

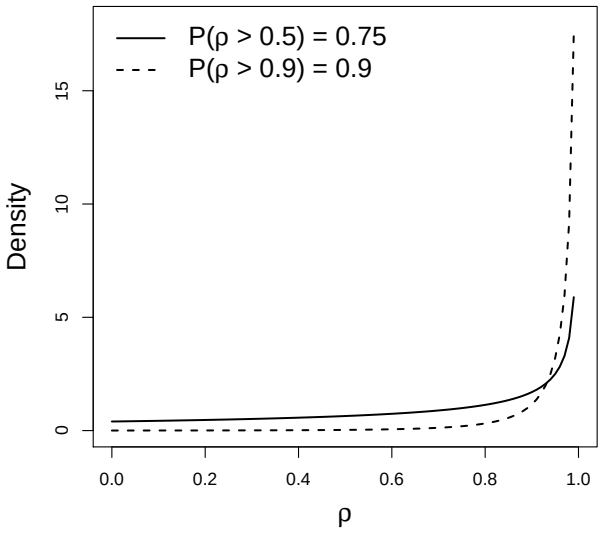

Figure 1: PC prior for ρ under the exchangeable model. The base model is $\rho = 1$.

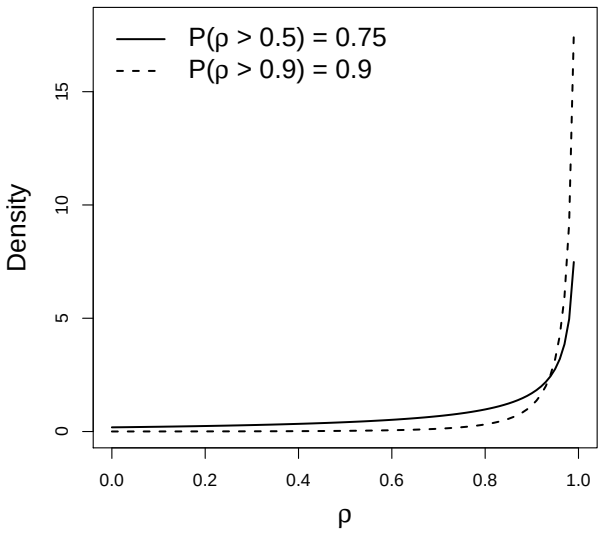

Figure 2: PC prior for ρ under the AR(1) model. The base model is $\rho = 1$.

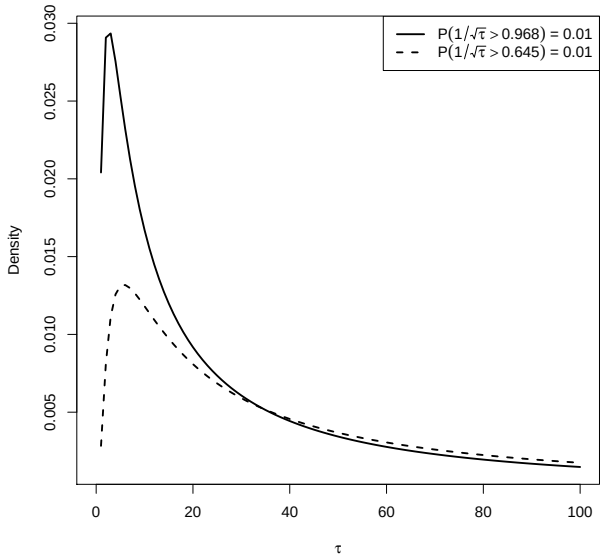

Figure 3: PC prior for τ under the RW model. The base model is $\tau = \infty$.

4.3 The structured case: spatial variation

4.3.1 Areal spatial variation

It is clear from Equation (8) that the ICAR model can be seen as a RW1 model (Equation (5) with $\texttt{rank}(\boldsymbol{R}) = n - 1$), and hence the PC prior for τ follows from Section 4.2.2.

4.3.2 Continuous spatial variation

PC priors for the range and marginal variance parameters of a GRF with *Matèrn* covariance function have been derived by Fuglstad et al. (2018). The joint PC prior for (τ, ϕ) with base model at $\tau = \infty$, $\phi = \infty$:

$$\pi(\tau, \phi) = \tilde{\lambda}_\phi \phi^{-2} \exp\left(-\tilde{\lambda}_\phi \phi^{-1}\right) \frac{\tilde{\lambda}_\tau}{2} \tau^{-3/2} \exp\left(-\frac{\tilde{\lambda}_\tau}{\sqrt{\tau}}\right) \qquad , \tau > 0, \phi > 0 \tag{14}$$

where, once the user fixes U_ϕ,a_ϕ,U_τ,a_τ such that $\mathbb{P}(\phi < U_\phi) = a_\phi$, $\mathbb{P}(1/\sqrt{\tau} > U_\tau) = a_\tau$ the parameters $\tilde{\lambda}_\phi$, $\tilde{\lambda}_\tau$ are calculated as

$$\tilde{\lambda}_\phi = -\log(a_\phi) U_\phi, \qquad \tilde{\lambda}_\tau = -\frac{\log(a_\tau)}{U_\tau}.$$

5 Examples

In the previous section we have shown how PC priors for varying coefficient models can be derived in a unified way. Here we illustrate their application in two spatial examples where varying coefficient models are relevant. All models are fitted within the `R-INLA` package (Martins et al., 2013) and the code is available in the supplementary material. The dataset used in example 5.2 is freely available, while the data from the example in Section 5.1 cannot be published due to privacy issues, but the related `R-INLA` code is available using a simulated similar dataset.

5.1 PM_{10} and hospital admissions in Torino, Italy

The goal is to estimate the effect of PM_{10} on the risk of hospitalization for respiratory causes using data on daily hospital admission from hospital discharge registers for the 315 municipalities in the province of Torino, Italy in 2004. In total, there are 12743 residents hospitalized for respiratory causes, aggregated by municipality and day. A reduced form of this dataset is available in the book by Blangiardo and Cameletti (2017). Daily average temperature (Kelvin degrees) and particular matter PM_{10} ($\mu g/m^3$) data are available at municipality level, the latter as estimates based on daily average PM_{10} concentration (Finazzi et al., 2013).

We consider the following model (all covariates are standardized):

$$
\begin{aligned}
y_{i,t} &\sim \text{Poisson}(E_{i,t}\exp(\eta_{i,t})) \\
\eta_{i,t} &= \alpha_t + u_i + \gamma \text{temp}_{i,t} + \beta_0 \text{PM}_{10,i,t} + \beta_i \text{PM}_{10,i,t} && (15) \\
(\alpha_1, ..., \alpha_{366})^{\mathsf{T}} &\sim \text{cyclic RW2}(\tau_{\text{rw2}}) && (16) \\
(u_1, ..., u_{315})^{\mathsf{T}} &\sim \text{BYM}(\tau_{\text{bym}}, \gamma_{\text{bym}}) && (17) \\
(\beta_1, \ldots, \beta_n)^{\mathsf{T}} &\sim \text{ICAR}(\tau_{\text{icar}}) && (18)
\end{aligned}
$$

where $y_{i,t}$ and $E_{i,t}$ are the observed and expected number of hospitalizations in municipality $i = 1, \ldots, 315$ and day $t = 1, \ldots, 366$ respectively and $\exp(\eta_{i,t})$ is the relative risk of hospitalization in municipality i and time t. Temperature (temp) is introduced as a fixed effect, as it is well known to be a confounder for the relationship between air pollution and health. $\text{PM}_{10,i,t}$ is taken as the sum of estimated daily average concentrations in the three days before t, in region i. The pollution effect is allowed to vary from municipality to municipality; we impose a sum to zero constraint on β_i to ensure identifiability of β_0, with $\beta_0 \sim N(0, 1000)$.

The random effects (16) and (17) capture residual temporal and spatial structure, respectively. The temporal random effects are assigned a RW2 wrapped on a circle to ensure a cyclic trend over time. The spatial random effect u_i is the sum of two random effects associated to municipality i, one spatially structured and one spatially unstructured, as defined by the popular BYM (Besag, York and Mollié) model (Besag et al., 1991). We follow the BYM parametrization introduced by Riebler et al. (2016) and use the PC priors derived therein for the two hyperparameters of the BYM: a marginal precision τ_{bym}, that allows shrinkage of the risk surface to a flat field, and a mixing parameter $\gamma_{\text{bym}} \in (0, 1)$, that handles the contribution from the structured and unstructured components. For ease of notation, in (17) we skip all the details and refer the reader to Riebler et al. (2016), formula (7).

Table 1 summarizes the selected U and a for all PC priors. We can use the practical rule of thumb described at the end of Section 4.2.2 to set an upper bound for the standard deviation. Weak prior knowledge suggests an upper bound for the marginal standard deviation approximately equal to $1, 3$ and 0.1 for the temporal trend (α_t), the spatial component (u_i) and the VC (β_i), respectively. For instance, the choice of $U = 0.1$ for β_i is to be interpreted as: there is roughly 95% probability that $\beta_i \in (e^{-0.1 \cdot 1.96}, e^{0.1 \cdot 1.96})$, i.e. there is little chance that the deviation in increased relative risk (associated to $1\mu g/m^3$ increase in PM_{10}) is larger than 1.2 in a given area.

The change in the posterior relative risk for a $10\mu g/m^3$ increase in PM_{10} is 1.002 (with 95% credible interval (0.998,1.006)). Figure 4 (panel a) displays the posterior mean for β_i, i.e. the municipality specific deviations (in the linear predictor scale) from the mean effect of PM_{10}. Panel (b) in Figure 4 shows the posterior probability of an increased risk associated to pollution, demonstrating that changes in the VCs across municipalities may only be substantial in the municipality of Turin (the *hotspot* in the south-east area). Looking at the prior vs posterior in Figure 5 (a), we see that there seems to be some information in the data regarding τ_{icar}.

From an epidemiological point of view, there seems to be two possible explanations for a spatially-varying pollution effect. First, the result might be due to the effect of an unobserved confounding variable which is not captured by the random effects in the model. Second, the PM_{10} chemical composition might change substantially over space, so that the PM_{10} may be more or less dangerous for people, according to where they live.

Sensitivity analysis

An interesting question is how sensitive the model fit is to a change in the PC prior parameters U, a. Figure 5(b) displays posterior distributions for τ_{icar} under three different settings (see Table 2) with increasing penalty for deviating from the base model. There does not seem to be a great effect of U on the posterior for τ_{icar} unless we impose a strong penalization for deviating from the base model (`pc3`). In terms of posterior relative risks, results (not reported here) remain basically unchanged across the different prior scenarios, unless a prior for the precision that puts a lot of probability mass around the base model is used, in which case the risk pattern is more shrunk towards no variation.

PC prior	α_t (rw2)	u_i (BYM)	β_i (ICAR)
$\pi(\tau \mid U, a = 0.01)$	$U = 0.1/0.31$	$U = 3/0.31$	$U = 0.1/0.31$
$\pi(\gamma \mid U, a = 0.5)$	-	$U = 0.5$	-

Table 1: Summary of the PC prior parameters U and a used in model (15) for the precisions (τ) and the γ parameter.

PC prior parameters	`pc1`	`pc2`	`pc3`
U	1/0.31	0.1/0.31	0.01/0.31
a	0.01	0.01	0.01

Table 2: Summary of the PC prior parameters U and a for τ_{icar} used in the sensitivity analysis for Model (15).

A possible alternative could be to assume a exchangeable model for the varying coefficient. Given the large number of areas ($n = 315$) we considered it was more natural to assume the varying coefficients to be spatially structured but for similar applications with a small number of areas an exchangeable model could be used.

5.2 House prices in Baton Rouge, Louisiana

The dataset considered in this example is available in Banerjee et al. (2015) and consists of selling prices ($) of 70 single family homes in East Baton Rouge Parish, Louisiana, sold in June 1989. Living area (square feet) and other area (square feet) such as garden, garage, etc., are available as covariates, as well as the longitude (lon) and latitude (lat) coordinates. An extended version of this dataset is analyzed in Gelfand et al. (2003). The spatial locations of the houses sold can be seen in Figure 6, along with the border delimiting the parish of East Baton Rouge. Even though the expectation is that bigger houses with a bigger external area are more expensive than smaller ones, location plays an important role in determining the price of a house. Hence, we allow for a spatially varying effect of living area (area) and other area (Oarea) in the following model (where the covariates have been standardized):

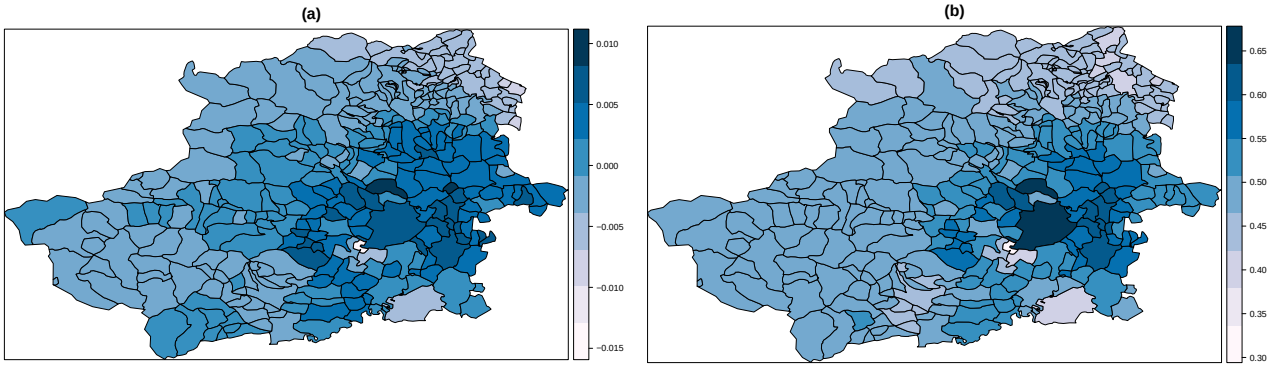

Figure 4: Posterior mean for the varying coefficients β_i (panel a) and posterior probability $\mathbb{P}(\beta_i > 0|\boldsymbol{y})$ (panel b).

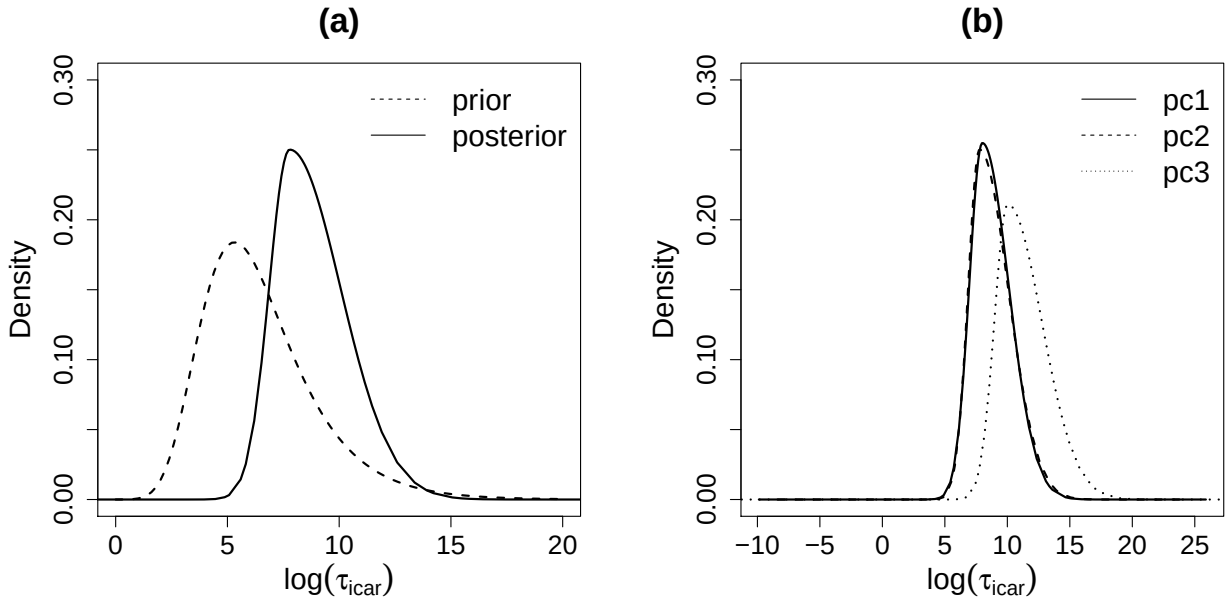

Figure 5: Prior *vs* posterior comparison for the precision parameter τ_{icar} as specified in Table 1 (panel a) and posterior for τ_{icar} for each setting in Table 2 (panel b).

$$
\begin{aligned}
\log(\text{price})_i &= \alpha + \gamma_{lon}\text{long}_i + \gamma_{lat}\text{lat}_i + \beta_{a,i}\text{area} + \beta_{b,i}\text{Oarea} + \epsilon_i + e_i && (19)\\
&(\beta_{a,1}, ..., \beta_{a,n})^{\mathsf{T}} \sim \mathcal{N}\left(\mathbf{0}, \tau_a^{-1}\boldsymbol{R}(\phi_a)\right) && (20)\\
&(\beta_{b,1}, ..., \beta_{b,n})^{\mathsf{T}} \sim \mathcal{N}\left(\mathbf{0}, \tau_b^{-1}\boldsymbol{R}(\phi_b)\right) && (21)\\
&\epsilon_i \sim \mathcal{N}(0, \tau_\epsilon^{-1}\boldsymbol{R}(\phi_\epsilon)) && (22)\\
&e_i \sim \mathcal{N}(0, \tau_e^{-1}) && (23)
\end{aligned}
$$

with $\boldsymbol{R}(\phi)$ as in Equation (9). PC priors for the parameters of the *Matèrn* covariance functions ϕ_a, τ_a, ϕ_b, τ_b and $\phi_\epsilon, \tau_\epsilon$ were scaled as follows. The maximum distance between observed locations is 5.12, so we set $U_\phi = 2$ and $a_\phi = 0.5$ so that $\mathbb{P}(\phi < 2) = 0.5$ for all ϕ_a, ϕ_b and ϕ_ϵ. Regarding the marginal standard deviation, prior knowledge on the scale of the response and of the covariates can be used to select U_τ and a_τ in a reasonable way; we set $U_\tau = 0.1/0.31$ and $a_\tau = 0.01$ for τ_a and τ_b (i.e. $\mathbb{P}(1/\sqrt{\tau} > 0.1/0.31) = 0.01$) and $U_\tau = 0.4/0.31$ and $a_\tau = 0.01$ for τ_ϵ (i.e. $\mathbb{P}(1/\sqrt{\tau} > 0.4/0.31) = 0.01$).

The posterior varying coefficient estimates for area and other area are shown in Figure 6. The effect of living area on log selling price (panel a) is greater than that of other area (panel b) and changes depending on location; in particular, there are two hot-spots where the effect appears to be greatest. The one on the left roughly corresponds to the area where Baton Rouge, capital of the state of Louisiana, is located. The bottom right corner corresponds to a district where household income is greater than that of the region as a whole.

The effect of other area on log selling price also varies spatially at it can be seen in Figure 6 (b). In particular, the red spot on the left hand side is roughly located on downtown Baton Rouge, the historic area of the city. On the other hand, it seems plausible that for houses located on the outskirts of the main cities in the region, the variable other area does not have such a strong impact on house price.

A small sensitivity analysis (see Table 3), was carried out in order to assess the impact of varying U and a. The results (not shown here) seldom vary unless a PC prior for τ with nearly all the mass concentrated on the base model (`pc.b`) is used (as already observed in Example 5.1). In practice, it is not possible to disentangle the effect of the range and marginal variance of a GRF. This results in sometimes different posterior means and distributions for the parameters under the remaining prior specifications in Table 3 but with essentially no differences in the fitted surfaces with respect to those shown in Figure 6. Given this difficulty in separating the effect of parameters ϕ and τ we opted to use an informative prior for the marginal variance, where U and a can be set in a more intuitive way, and a less informative prior for the range parameter.

	scenario									
	a_i		pc.a		pc.b		pc.c		pc.d	
	a_ϕ	a_τ	U_ϕ	U_τ	U_ϕ	U_τ	U_ϕ	U_τ	U_ϕ	U_τ
$\beta_{a,i}$	0.5	0.01	2	1/0.31	2	0.01/0.31	0.5	0.1/0.31	5	0.1/0.31
$\beta_{b,i}$	0.5	0.01	2	1/0.31	2	0.01/0.31	0.5	0.1/0.31	5	0.1/0.31
ϵ_i	0.5	0.01	2	4/0.31	2	0.04/0.31	0.5	0.4/0.31	5	0.4/0.31

Table 3: Summary of the PC prior parameters U and a used in model (19) for the precisions (τ) and the ϕ parameters in the sensitivity analysis.

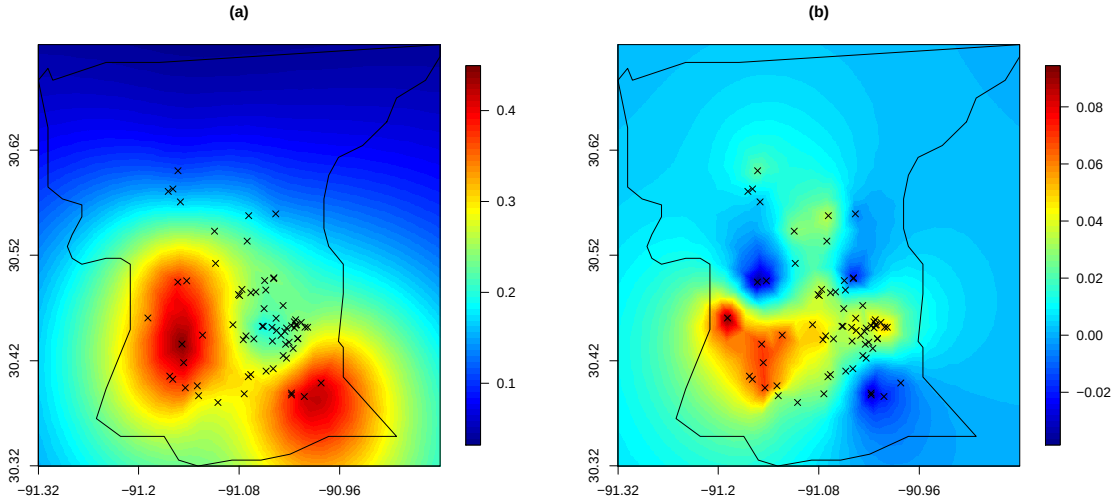

Figure 6: Posterior mean for the varying coefficient of area $\beta_{\mathrm{a},i}$ (panel a) and other area $\beta_{\mathrm{b},i}$ (panel b). Observed locations are marked with a cross.

6 Discussion

This paper presents varying coefficient models as a single class of models defined by Gaussian processes that only differ in the structure of the covariance matrix. This means that the various models considered can be treated similarly. Further, we present a unified approach for setting hyperpriors for varying coefficient models, regardless of the model assumed on the coefficients. The definition of the varying coefficient model as a flexible extension of a simpler model *calls for* eliciting priors that allow the simpler base model to arise. PC priors guarantee this; since the mode is at the base model, overfitting, a common aspect in complex hierarchical models, is avoided by construction.

We have illustrated the use of PC priors for varying coefficients in two different applications. Whether the covariate is standardized or not obviously makes an impact on the scale of the varying coefficient, thus the user should be careful in defining the value U for the precision parameter τ and change it accordingly if the scale of the covariate is transformed. In general the choice of U does not impact much the posterior for $\boldsymbol{\beta}$, unless almost all the probability mass is assigned deliberately to the base model, i.e. unless an unreasonable prior is used, meaning a prior that is against our prior knowledge on the behaviour of the parameter. Building a prior on the distance from a base model allows the level of informativeness of the prior to be set according to the actual amount of prior information. In the VCM case, for instance, the PC prior can be set as a weakly informative prior for the precision as we usually have a reasonable guess on the scale of the varying coefficient (depending on the link function of the model, the scale of the data and of the covariate). On the contrary, the PC prior can act as a vague prior for the correlation parameter, by just setting $U = a = 0.5$ in order to express ignorance.

Choice of the prior $\pi(\xi)$ is difficult in practice, because there is typically no prior information on the hyperparameters in hierarchical models. Moreover, the empirical information available to estimate the posterior for ξ is less compared to that available for the parameters in the linear predictor. This means that the the prior for ξ is deemed to have a large impact on the model, especially in sparse data cases. In our opinion, this represents a further good reason for using PC priors, as we can be more confident that no overfitting takes place when there is not enough information in the data. Even though we do not know much at prior about suitable values for ξ, we often know exactly what

a hyperparameter *does* in terms of shrinkage to a simpler model. Working with priors that allow to introduce information regarding what the parameters mean can help the user to choose sensible values for the prior.

Acknowledgements

Maria Franco-Villoria and Massimo Ventrucci are supported by the PRIN 2015 grant project n.20154X8K23 (EPHASTAT) founded by the Italian Ministry for Education, University and Research.

References

Banerjee, S., Carlin, B., and Gelfand, A. (2015). *Hierarchical Modeling and Analysis for Spatial Data, Second Edition*. CRC Press/Chapman & Hall. Monographs on Statistics and Applied Probability.

Besag, J. (1974). Spatial interaction and the statistical analysis of lattice systems (with discussion). *Journal of the Royal Statistical Society Series B*, 36(2):192–225.

Besag, J., York, J., and Mollie, A. (1991). Bayesian image restoration, with two applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, 43:1–21.

Blangiardo, M. and Cameletti, M. (2017). *Spatial and Spatio-temporal Bayesian Models with R-INLA*. Wiley.

Cai, Z. and Sun, Y. (2003). Local linear estimation for time-dependent coefficients in Cox's regression models. *Scandinavian Journal of Statistics*, 30:93–11.

Fan, J. and Zhang, W. (1999). Statistical estimation in varying coefficient models. *The Annals of Statistics*, 27:1491–1518.

Fan, J. and Zhang, W. (2008). Statistical methods with varying coefficient models. *Statistics and Its Interface*, 1:179–195.

Ferguson, C., Bowman, A., Scott, E., and Carvalho, L. (2007). Model comparison for a complex ecological system. *Journal of the Royal Statistical Society Series A*, 170(3):691–711.

Finazzi, F., Scott, M., and Fasso, A. (2013). A model-based framework for air quality indices and population risk evaluation, with an application to the analysis of Scottish air quality data. *Journal of the Royal Statistical Society Series C*, 62(2):287–308.

Finley, A. (2011). Comparing spatially-varying coefficients models for analysis of ecological data with non-stationary and anisotropic residual dependence. *Methods in Ecology and Evolution*, 2:143–154.

Frühwirth-Schnatter, S. and Wagner, H. (2010). Stochastic model specification search for Gaussian and partial non-Gaussian state space models. *Journal of Econometrics*, 154(1):85–100.

Frühwirth-Schnatter, S. and Wagner, H. (2011). Bayesian variable selection for random intercept modeling of Gaussian and non-Gaussian data. In *J. M. Bernardo, M. J. Bayarri, J. O. Berger, A. P. Dawid, D. Heckerman, A. F. M. Smith and M. West (Eds.)*, pages 165–200. Bayesian Statistics 9, Oxford.

Fuglstad, G. A., Simpson, D., Lindgren, F., and Rue, H. (2018). Constructing priors that penalize the complexity of Gaussian random fields. *Journal of the American Statistical Association*.

Gelfand, A., Kim, J., Sirmans, C., and Banerjee, S. (2003). Spatial modeling with spatially varying coefficient processes. *Journal of the American Statistical Association*, 98(462):387–396.

Hastie, T. and Tibshirani, R. (1993). Varying-coefficient models. *Journal of the Royal Statistical Society Series B*, 55(4):757–796.

Hoover, D., Rice, J., and Wu, C. (1998). Nonparametric smoothing estimates of time-varying coefficient models with longitudinal data. *Biometrika*, 85(4):809–822.

Huang, J., Wu, C., and Zhou, L. (2002). Varying-coefficient models and basis function approximations for the analysis of repeated measurements. *Biometrika*, 89(1):111–128.

Kimeldorf, G. and Wahba, G. (1970). A correspondence between Bayesian estimation on stochastic processes and smoothing by splines. *The Annals of Mathematical Statistics*, 41(2):495–502.

Klein, N. and Kneib, T. (2016). Scale-dependent priors for variance parameters in structured additive distributional regression. *Bayesian Analysis*, 11(4):1071–1106.

Kullback, S. and Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, 22:79–86.

Laird, N. and Ware, J. (1982). Random-effects models for longitudinal data. *Biometrics*, 38(4):963–974.

Lindgren, F. and Rue, H. (2008). On the second-order random walk model for irregular locations. *Scandinavian Journal of Statistics*, 35(4):691–700.

Martins, T. G., Simpson, D., Lindgren, F., and Rue, H. (2013). Bayesian computing with INLA: New features. *Computational Statistics & Data Analysis*, 67(0):68–83.

Marx, B. (2010). P-spline varying coefficient models for complex data. In *T.Kneib and G. Tutz (Eds.)*. Statistical Modelling and Regression Structures, Physica-Verlag HD.

Mu, J., Wang, G., and Wang, L. (2018). Estimation and inference in spatially varying coefficient models. *Environmetrics*, 29.

Park, B., Mammen, E., Lee, Y., and Lee, E. (2015). Varying coefficient regression models: A review and new developments. *International Statistics Review*.

Riebler, A., Sørbye, S. H., Simpson, D., and Rue, H. (2016). An intuitive Bayesian spatial model for disease mapping that accounts for scaling. *Statistical Methods in Medical Research*, 25(4):1145–1165. PMID: 27566770.

Rue, H. and Held, L. (2005). *Gaussian Markov Random Fields*. Chapman and Hall/CRC.

Simpson, D., Rue, H., Riebler, A., Martins, T. G., and Sørbye, S. H. (2017). Penalising model component complexity: A principled, practical approach to constructing priors. *Statistical Science*, 32(1):1–28.

Sørbye, S. and Rue, H. (2014). Scaling intrinsic Gaussian Markov random field priors in spatial modelling. *Spatial Statistics*, 8:39–51.

Sørbye, S. and Rue, H. (2017). Penalised complexity priors for stationary autoregressive processes. *Journal of Time Series Analysis*, 38:923–935.

Staubach, C., Schmid, V., Knorr-Held, L., and Ziller, M. (2002). A Bayesian model for spatial wildlife disease prevalence data. *Preventive Veterinary Medicine*, 56:75–87.

Stein, M. (1999). *Interpolation of Spatial Data: Some Theory for Kriging*. Springer-Verlag, New York.

Tian, L., Zucker, D., and Wei, L. (2005). On the Cox model with time-varying regression coefficients. *Journal of the American Statistical Association*, 100(469):172–183.

Waller, L., Zhu, L., Gotway, C., Gorman, D., and Gruenewald, P. (2007). Quantifying geographic variations in associations between alcohol distribution and violence: a comparison of geographically weighted regression and spatially varying coefficient models. *Stochastic Environmental Research and Risk Assessment*, 21:573–588.

Warnes, J. and Ripley, B. (1987). Problems with likelihood estimation of covariance functions of spatial Gaussian processes. *Biometrika*, 74(3):640–642.

Zhang, H. (2004). Inconsistent estimation and asymptotically equal interpolations in model-based geostatistics. *Journal of the American Statistical Association*, 99(465):250–261.

A Appendix: Derivation of the PC prior

A.1 The unstructured case

The varying coefficient model in the exchangeable case is

$$\begin{array}{ll}\eta_t = \alpha + \beta_t x_t & t = 1, ..., n, \\ \boldsymbol{\beta} \sim \mathcal{N}(0, \boldsymbol{R}(\rho)), & \end{array}$$

with

$$\boldsymbol{R}(\rho) = \begin{bmatrix} 1 & \rho & \dots & & \rho \\ \rho & 1 & \rho & \dots & \rho \\ \cdot & & \cdot & & \cdot \\ \cdot & & & \cdot & \cdot \\ \rho & \rho & \dots & \rho & 1 \end{bmatrix}$$

and base model $\rho = 1$ (i.e. $\beta_t = \beta \ \ \forall t$). To evaluate the distance from the base model we need to use a limiting argument. For a fixed value of ρ_0 close to 1, the KLD distance:

$$KLD(f_1(\rho)||f_0) = \frac{1}{2}\left(\frac{(n-1)(1-\rho)}{1-\rho_0} - n - \log(1+(n-1)\rho) + \log n - (n-1)\log\frac{1-\rho}{1-\rho_0}\right)$$

Considering the limiting value as $\rho_0 \to 1$, the distance

$$d(\rho) = \lim_{\rho_0 \to 1} \sqrt{2KLD(f_1(\rho)||f_0)} = \lim_{\rho_0 \to 1} \sqrt{\frac{(n-1)(1-\rho)}{1-\rho_0}} = c\sqrt{1-\rho}, \quad 0 \le \rho \le 1$$

for a constant c that does not depend on ρ. Since $0 \le d(\rho) \le c$, assigning a truncated exponential with rate λ on $d(\rho)$ we have

$$\pi(d(\rho)) = \frac{\lambda \exp(-\lambda c\sqrt{1-\rho})}{1-\exp(-\lambda c)}, \qquad 0 \le d(\rho) \le c, \quad \lambda > 0.$$

Reparametrizing $\theta = \lambda c$ leads to the PC prior for ρ:

$$\pi(\rho) = \frac{\theta \exp(-\theta\sqrt{1-\rho})}{2\sqrt{1-\rho}(1-\exp(-\theta))}, \quad 0 \le \rho \le 1, \quad \theta > 0.$$

A.2 The autoregressive model of first order

The varying coefficient model in the AR1 case is

$$\begin{array}{ll}\eta_t = \alpha + \beta_t x_t & t = 1, ..., n, \\ \boldsymbol{\beta} \sim \mathcal{N}(0, \boldsymbol{R}(\rho)), & \end{array}$$

with $\boldsymbol{R}(\rho)_{ij} = (\rho^{|i-j|})$ and base model $\rho = 1$. Using a limiting argument similar to that of Appendix A.1, the distance to the base model is

$$d(\rho) = c\sqrt{1-\rho}, \qquad |\rho| < 1 \tag{24}$$

where c is a constant. Note that (24) is upper bounded, $0 \le d(\rho) \le c\sqrt{2}$, so that the PC prior for $d(\rho)$ is

$$\pi(d(\rho)) = \frac{\lambda \exp(-\lambda c\sqrt{1-\rho})}{1-\exp(-\lambda c\sqrt{2})}, \qquad 0 \le d(\rho) \le c\sqrt{2}, \quad \lambda > 0.$$

Reparametrizing $\lambda = \theta/c$ and using the change of variable formula it follows that the PC prior on the ρ scale is (Sørbye and Rue, 2017)

$$\pi(\rho) = \frac{\theta \exp(-\theta\sqrt{1-\rho})}{(1-\exp(-\sqrt{2}\theta))2\sqrt{1-\rho}}, \qquad |\rho| < 1, \quad \theta > 0.$$

A.3 Random walk model of order one and two

The varying coefficient has a joint distribution given by

$$\boldsymbol{\beta} \sim \mathcal{N}(\mathbf{0}, \tau^{-1}\boldsymbol{R}^{-1})$$

with $\boldsymbol{R}$ symmetric semi-positive definite matrix. Let $f_0 = \pi(\boldsymbol{\beta}|\tau_0 = \infty)$ and $f = \pi(\boldsymbol{\beta}|\tau)$ denote the base and flexible models, with precisions τ_0 and τ, respectively. Simpson et al. (2017) show that $\text{KLD}(f||f_0)$ goes to $\frac{\tau_0 n}{2\tau}$, for τ much lower than τ_0 and $\tau_0 \to \infty$, so that $d(\tau) = \sqrt{2\text{KLD}(f||f_0)} = \sqrt{\tau_0 n/\tau}$ and $d(\tau) \sim \exp(\lambda)$, $\lambda > 0$.

By a change of variable and setting the rate $\lambda = \theta/\sqrt{n\tau_0}$, Simpson et al. (2017) derive the PC prior for τ as

$$\pi(\tau) = \frac{\theta}{2}\tau^{-3/2}\exp\left(-\theta/\sqrt{\tau}\right), \quad \tau > 0, \theta > 0, \tag{25}$$

which is a Gumbel$(1/2, \theta)$ type 2 distribution.

A.4 Continuous spatial variation

The spatially varying coefficient can be seen as a realization of a Gaussian random field (GRF)

$$\boldsymbol{\beta} \sim \mathcal{N}\left(\mathbf{0}, \tau^{-1}\boldsymbol{R}(\phi)\right)$$

with *Matèrn* correlation function as in (14). PC priors for the range and marginal variance parameters of a GRF with *Matèrn* covariance function have been derived by Fuglstad et al. (2018). Here we only summarize the main results on the computation of the PC prior, while for further details the reader is referred to Fuglstad et al. (2018). Deriving PC priors for these parameters is more complex that in the previous situations considered in this paper due to the infinite-dimensional nature of GRFs. Following Fuglstad et al. (2018) and setting $d = 2$, parameters ϕ and τ are conveniently reparametrized as:

$$\kappa = \frac{\sqrt{8\nu}}{\phi} \qquad \psi = \sqrt{\tau^{-1}}\phi^{\nu}\sqrt{\frac{\Gamma(\nu+1)(4\pi)}{\Gamma(\nu)}}$$

Since the parameter ψ depends on κ, the joint PC prior is built as $\pi(\psi, \kappa) = \pi(\kappa)\pi(\psi|\kappa)$, which can then be transformed into a joint PC prior for (ϕ, τ). In this case, the base model corresponds to $\phi = \infty$ (or equivalently, $\kappa = 0$), i.e. the spatial correlation is so strong that we have a constant field and $\tau = \infty$ ($\psi = 0$), i.e. no marginal variance. The PC prior $\pi(\psi|\kappa)$ is built based on the observations available at n locations, while the PC prior $\pi(\kappa)$ is based on the infinite-dimensional GRF to avoid a model-dependent prior; see Fuglstad et al. (2018) for details.

The PC prior for κ:

$$\pi(\kappa) = \lambda_1 \exp\left(-\lambda_1\kappa\right), \quad \kappa > 0, \tag{26}$$

and $\lambda_1 > 0$. The user can set U_1 and a_1 such that $\mathbb{P}(\phi < U_1) = a_1$, so that $\lambda_1 = -\left(\frac{U_1}{\sqrt{8\nu}}\right)\log(a_1)$.

The PC prior for $\psi|\kappa$ follows an exponential distribution:

$$\pi(\psi|\kappa) = \lambda_2 \exp(-\lambda_2 \psi), \quad \psi > 0 \tag{27}$$

where, as before, $\lambda_2 > 0$ can be selected based on the user-selected values U_2 and a_2 such that $\mathbb{P}(1/\sqrt{\tau} > U_2|\kappa) = a_2$, which leads to $\lambda_2(\kappa) = -\kappa^{-\nu}\sqrt{\frac{\Gamma(\nu)}{\Gamma(\nu+1)(4\pi)}}\frac{\log(a_2)}{U_2}$.

The joint PC prior $\pi(\kappa, \psi) = \pi(\kappa)\pi(\psi|\kappa)$, and by a change of variable it follows that the PC prior for τ, ϕ:

$$\pi(\tau, \phi) = \tilde{\lambda}_\phi \phi^{-2} \exp\left(-\tilde{\lambda}_\phi \phi^{-1}\right) \frac{\tilde{\lambda}_\tau}{2} \tau^{-3/2} \exp\left(-\frac{\tilde{\lambda}_\tau}{\sqrt{\tau}}\right) \qquad , \tau > 0, \rho > 0 \tag{28}$$

where, once the user fixes U_ϕ,a_ϕ,U_τ,a_τ such that $\mathbb{P}(\phi < U_\phi) = a_\rho$, $\mathbb{P}(1/\sqrt{\tau} > U_\tau) = a_\tau$ the parameters $\tilde{\lambda}_\phi$, $\tilde{\lambda}_\tau$ are calculated as

$$\tilde{\lambda}_\phi = -\log(a_\phi) U_\phi, \qquad \tilde{\lambda}_\tau = -\frac{\log(a_\tau)}{U_\tau}.$$